OPINION

Error prevention and mitigation as forces in the evolution of genes and genomes

Tobias Warnecke and Laurence D Hurst

Abstract | Why are short introns rarely a multiple of three nucleotides long? Why do essential genes cluster? Why are genes in operons often lined up in the order in which they are needed in the encoded pathway? In this Opinion, we argue that these and many other – ostensibly disparate – observations are all pieces of an emerging picture in which multiple aspects of gene anatomy and genome architecture have evolved in response to error-prone gene expression.

Subject categories: Evolutionary Biology, Gene Expression

Faithful information transmission is critical to life. This is perhaps most evident when genetic instructions are passed on from parents to their offspring. If information is corrupted at this stage, the result – depending on the sequence affected – may be lethal. A complex suite of mechanisms is in place to avoid detrimental effects of this kind, starting from the inbuilt capacity of DNA polymerases to backtrack and proof-read their own output. Fidelity of information transmission is not only critical during replication, however, but also during the day-to-day running of the cell. Information encoded in the DNA must be read off and processed to yield biological effector molecules. This is typically a multi-step process, providing ample opportunity for errors to be made along the way. A gene can be transcribed at the wrong time or at levels too high or too low to achieve optimal functioning; it can be mis-transcribed, mis-spliced or mis-translated. Once translated, the protein can misfold, localize incorrectly, or be activated or degraded too soon or too late.

Safeguarding the integrity of biological information, the cellular machines that decode that information (such as the ribosome and the spliceosome) operate with intrinsically high fidelity[1] and there exist a plethora of quality control pathways that detect and eliminate erroneous gene products. Exactly how many errors are made and caught after the act, however, also depends on the gene that is being expressed. Some transcripts are particularly susceptible to accidental frame-shifts during translation; others habitually escape quality control so that errors may go unnoticed.

In this Opinion, we highlight how both gene anatomy (i.e. the composition and structure of genes and their products) and genome architecture (i.e. the arrangement of genes in the genome) have evolved to reduce the rate at which errors occur (error prevention) or curb deleterious effects if an error has already been made (error mitigation, Figure 1). Drawing on case studies from, amongst other systems, *E.coli*,

human, *S. cerevisiae*, and *Paramecium tetraurelia*, we first discuss how selection has moulded gene anatomies to facilitate high-fidelity information transmission or enable faulty products to be intercepted by quality control. Thereafter, we argue that non-random genome architecture – from the composition of local gene neighbourhoods to the differential distribution of genes across chromosomes – may also frequently reflect selection against erroneous expression. We focus specifically on how cells prevent transcript levels from falling below a critical threshold in the face of stochastic gene expression. Our aim here is not to present a comprehensive inventory of genomic adaptations to erroneous gene expression, We largely do not discuss adaptations relating to protein stability and misfolding, which have been well reviewed recently[2]. Neither do we address the creative potential of error-prone expression, where a lack of fidelity in gene expression generates natural variation that can result in increased fitness in novel environments[2]. Rather, we focus on a few examples that illustrate the diversity of molecular signatures associated with error management, whilst highlighting current progress and areas for future research.

The role of gene anatomy

Preventing faulty gene products. Arguably the most extensive support for a role of error prevention in the evolution of gene anatomy comes from the study of synonymous codon usage. Codons – both individually and in the context of their neighbours – differ in their propensity to be mistranslated or induce frame-shifts. In line with selection to reduce errors during translation, functionally important and structurally sensitive sites are enriched for less error-prone codons in a taxonomically diverse range of species (reviewed in Ref. 2). Beyond individual codons, certain codon combinations also appear to be avoided. Notably, protein-coding sequence in *S. cerevisiae*, *E. coli*, and *C. elegans* is depleted for mononucleotide repeats[3]. The same signature is absent in introns, supporting selective avoidance over mutational bias. As mononucleotide repeats are prone to slippage during transcription[4] and translation[5], the most parsimonious explanation is selection against error-prone nucleotide composition. However, whether selection principally operates to lower transcription or translation errors remains unclear.

Dealing with faulty gene products. Even if an error fails to be prevented, the fitness consequences that ensue may be minimal. Famously, neighbouring triplets in the genetic code tend to specify biochemically similar amino acids, so that single-nucleotide substitutions rarely lead to radical amino acid replacements[6]. This property – which may reflect past selection for an error mitigation capacity or constitute a byproduct of genetic code evolution[7] – makes the code robust to genetic mutations but also to transcriptional and translational errors[8], where misreading often affects a single base.

However, not all single-nucleotide changes have a minor impact on functionality. Notably, errors that create premature termination codons (PTCs) are unlikely to yield functional products. Translation of PTC-containing transcripts can have serious repercussions for fitness, wasting translational resources as

well as generating potentially toxic truncated peptides[9]. In eukaryotes, deleterious effects of PTC-containing transcripts are mitigated by the intervention of a dedicated quality control system: During the pioneer round of translation the nonsense-mediated decay (NMD) machinery recognizes that the PTC occurs too early (relative to a specific downstream features that differs across species) and triggers degradation of the transcript[10].

Intriguingly, the presence of NMD appears to have systematically affected the evolution of gene anatomies. The most striking evidence in this regard comes from the ciliate *Paramecium tetraurelia*. The majority (>96%) of introns in its genome are very short (<34nt) and, transcriptome analysis revealed high rates (~1%) of intron retention[11]. Curiously, fewer of these short introns than expected have a length divisible by three nucleotides (*3n*) and those that do are more likely to harbour in-frame stop codons[11]. Why would this be? Introns that are not *3n* lead to frame-shifts when accidentally retained in the mRNA, which, in turn, is likely to generate a PTC in the new frame[12] and render erroneous transcripts subject to NMD. In contrast, introns that neither contain in-frame PTCs nor cause PTCs by inducing a frame-shift escape detection by NMD and may be repeatedly translated into nonsensical protein. The scarcity of stopless *3n* introns, observed not only in *Paramecium* but also in species as diverse as humans, *Arabidopsis* and the fungus *Yarrowia lipolytica*[11, 13], therefore strongly suggests that intron length and composition have been shaped by selection to ensure that mis-spliced transcripts are recognized by NMD.

Some additionally suggest that selection has favoured the retention of nucleotide triplets that encode out-of-frame stops ("ambush codons") and therefore terminate translation when a frame-shift occurs upstream[14]. However, unless GC content is very high[15, 16], translation is terminated quickly after an accidental frame-shift anyway. This is because, at least in the human genome, common codons are often codons that generate a partial stop when a 1-base pair frameshift occurs in the 5' or 3' direction (Figure 2). In humans, on average only ~15 amino acids are translated before the ribosome encounters a stop in the new frame[12]. As marginal costs savings through dedicated ambush codons would therefore typically be minute, it remains questionable how many, if any, off-frame termination triplets are actually maintained by selection to provide ambush functionality.

Some classes of transcripts are unable to trigger NMD when a PTC is introduced and seem to have evolved alternative features to reduce the fitness cost of faulty products. In mammals, to be recognized by the major NMD pathway the PTC must be located some distance upstream of the last exon-junction complex, which is deposited during splicing. Intronless transcripts, therefore, have a problem: when a PTC is generated, for example during transcription, these transcripts cannot trigger NMD. Do these genes therefore have to bear a higher error load? Recent evidence suggests not. Intronless genes in mammals instead employ fewer codons for which the introduction of a single incorrect nucleotide during transcription results in a stop codon[17]. Thus, where errors cannot be mitigated one way, their impact may

be alleviated via a complementary route or selection may operate to reduce their incidence by promoting the fixation of less error-prone states.

Following the same logic, *E. coli* genes differ in their propensity to use translationally optimal codons depending on whether or not they are clients of the chaperonin GroEL, with obligate clients relatively depleted of optimal codons[18]. This is consistent with obligate GroEL substrates experiencing selective relief because the chaperone can mitigate at least some deleterious effects of mistranslation-induced misfolding errors.

Unraveling error-adapted gene anatomy. Dissecting differential interactions between individual genes and quality control machineries arguably constitutes the most informative route to understanding adaptive gene anatomy because it can reveal subsets of substrates that behave unlike others and can form the cornerstone for critical tests. Importantly, the need for studying interactions goes beyond considering core expression/quality control machineries. Ancillary interactions, those not directly involved in generating the error, nonetheless shape error propensities and hence the need for adaptive solutions. Consider transcripts populated by RNA-binding proteins, such as those deposited during splicing. Some of these transcripts may – based on sequence information alone – appear liable to erroneous interactions but, in fact, be protected by their binding partners. It appears, for instance, that a large number of cryptic polyadenylation sites in humans can persist without deleterious consequences because they are rendered unusable by nearby binding of U1 snRNP, which prevents cleavage[19]. Integrating increasingly comprehensive interaction data sets to capture the complex context of gene expression will therefore be essential for understanding adaptive gene anatomy.

The known unknowns. In both humans and *Paramecium* the proportion of transcripts estimated to contain splicing errors is comparatively high (with lower-bound estimates in the region of 1% [11, 20]). Similarly high or higher error rates have been reported for other steps of transcript processing, which might be indicative of efficient downstream error mitigation. In yeast, the fraction of transcripts that are polyadenylated prematurely, and therefore lack a termination signal, may be as high as 10%[21]. In the absence of a proper termination signal, attempts by the ribosome to translate the polyA tail lead to mRNA degradation and translational repression. This suggests that polyA tails function as part of an error control system[22, 23]. Yet most incorrectly polyadenylated transcripts never reach the ribosome, being degraded by the nuclear exosome at the site of transcription[24]. Despite the high apparent error rate, we know little about the sequence features involved in this mitigation process. Targeted knock-down of exosome components in conjunction with high-throughput sequencing might shed light on this issue.

For yet other stages of transcript processing even basic error estimates are lacking. For example, how many phosphorylation/dephosphorylation events happen off-target or at the wrong time? Importantly, this is not simply a problem of quantification. The hard problem is telling errors from functional isoforms (Box 1), especially when (ostensible) telltale signs like PTCs are absent or unknown.

The role of genome architecture

Genome architecture, i.e. the order, spacing, and orientation of genes in the genome, can be highly non-random[25]. In part, this reflects the action of selection. Genes, through recombination, retroposition or similar processes, repeatedly sample genomic real estate and – over evolutionary time – come to reside in locations that confer high fitness. In bacteria, non-random gene order is primarily owing to the fact that gene expression is organized into polycistronic transcripts, with genes participating in the same biochemical pathway or protein complex often co-located in the same operon. However, pathway-based clustering of genes cannot explain every aspect of non-random genome architecture. For example, in *E. coli* and other bacteria essential genes cluster around the origin of replication and preferentially reside on the leading strand[26]. While proximity to the origin may be adaptive because it enhances expression of core genes during multiple concurrent rounds of replication, preferential location on the leading strand is considered beneficial because it prevents the transcribing RNA polymerase from colliding head-on with the DNA polymerase during replication.

Controlling stochasticity. Ensuring production of transcripts does not suddenly cease or fluctuate violently is not only a challenge during cell division when the replication machinery competes for access to the DNA. Cells regularly need to adjust the expression of some genes without disturbing the expression of others, especially those sensitive to changes in dosage. In addition, molecular binding dynamics such as between transcription factors (or chromatin remodeling complexes) and DNA are intrinsically stochastic[27]. As a result, gene product levels may fall below (or, indeed, rise above) a critical threshold. The degree of stochasticity (noise) exhibited by individual genes, however, can vary dramatically[28], highlighting the possibility that noise is an evolvable trait[27, 29]. Stochasticity can be reduced by making use of specific promoter architectures (prominently, the absence of TATA motifs is associated with low-noise genes[30, 31]), raising overall expression levels[28], increasing gene copy number[27, 32], and altering genetic network wiring to include noise-abating feedback loops[33, 34]. In the remainder of this section, we argue that, in addition, genome architecture has, at several scales of organization, been moulded by selection to dampen noise.

Noise-abating genome architecture. Genes with similar noise tolerance are not randomly scattered across chromosomes but instead form clusters. Notably, in yeast, noise-sensitive genes (both essential and non-essential) cluster together[35]. These clusters lie in domains of open chromatin, suggesting that noisiness of individual genes is, at least in part, determined by regional chromatin states. This is in line with observations that, in both mammals and yeast, neighbouring genes exhibit correlated bursting kinetics and transgenes adopt the bursting kinetics of their new host domain[32, 36]. Indeed, a model in which genes are

allowed to recombine into domains of differential noise recreates the observed clustering of essential genes in 'quiet' domains[35].

Selection to dampen noise may also drive gene order and orientation at even finer levels. Pairs of genes that lie in head-to-head orientation - that is, those that are transcribed divergently, from the same promoter but from different strands - exhibit reduced noise[30, 37]. This is probably because the use of a shared bidirectional promoter generates a mutually reinforcing chromatin micro-environment that leads to reduced stochastic fluctuations[37]. Similarly, leaky transcription can be modulated by the expression of antisense transcripts, which often share their promoter with a downstream sense gene[37, 38]. Consistent with expectations, this type of organization is more common for genes expected to be more sensitive to noise, such as essential genes and genes that participate in protein complexes[37].

The fine-scale organization of some bacterial operons is also consistent with selection for noise abatement. Lovdok *et al.* compared operon structures across bacteria and found that some local gene arrangements (consisting of gene couples) in operons encoding the chemotaxis pathway are much more highly conserved than others. This is surprising: genes residing in the same operon are transcriptionally coupled, so why should there be selection to maintain a particular order? Intriguingly, the affected gene pairs exhibit strong coupling at the level of translation in *E. coli*, which the authors show reduces noise in the output generated by the pathway[39]. It therefore appears likely that these 'neighbourhoods' are preferentially conserved because they buffer pathway output against fluctuations in the concentration of individual proteins.

Noise also appears to have influenced gene order in some metabolic operons where, curiously, genes often lie in the order in which they are required in the corresponding pathway (a phenomenon dubbed colinearity[40]). Again, this is surprising given that the genes are transcriptionally coupled. However, it is consistent with a model where, at low transcription rates, pathways occasionally collapse (because – owing to stochastic effects - a critical component of the pathway fails to be present altogether) but are then more easily restarted when gene order is colinear[40]. Colinearity is indeed exclusive to lowly expressed operons[40].

Does X-treme noise foster relocation?

In *C. elegans*, where genome-wide RNAi knock-down data are available, only 5.6% of X-linked genes are essential compared to 12.8% on the autosomes[41]. X chromosomes in mammals and *C. elegans* are also depleted for genes shown to be haploinsufficient in yeast, although the same is not true for *Drosophila*[42]. Why might essential genes avoid X chromosomes? The answer may, as above, in part be related to noise.

Genes that are haploid regarding the number of chromosomal copies from which they are expressed (haploid-expressed genes) are expected to be high-noise genes[43]. This is because stochastic fluctuations in transcript production are more effectively dampened if a second target for transcription is present. As predicted, haploid-expressed human autosomal genes appear to be especially noisy[44]. In mammals, both sexes are effectively haploid for the X chromosome, either by virtue of being male or following inactivation of one copy in females. As essential and haploinsufficient genes tend to be low-noise genes[35], a simple rationalization is thus that essential genes are selected to avoid the high noise context of the haploid-expressed X.

This hypothesis is very much speculation. More stringent support is needed to reinforce the hypothesis that noise has contributed to inter-chromosomal differences in gene content and that higher ploidy, as predicted by theory[29], confers fitness benefits by reducing noise. One useful experiment would be to induce ploidy differences *de novo* in a suitable organism and subsequently assay noise. If the above hypothesis is correct, we would expect polyploidization to be associated with reduced within-gene variability in gene expression. The observation that in plants of the genus *Senecio* polyploidization of artificial hybrids globally reduces between-gene variance[45] is intriguing in this regard.

Conclusions

We have argued that multiple facets of gene anatomy and genome architecture may be adaptations to error-prone gene expression. Concerning genome architecture**,** the evolution of non-random gene neighbourhoods might often reflect selection to prevent detrimental stochastic fluctuations in transcript levels. Current evidence remains largely limited to a handful of model species, however, so that, in order to assess the relative role of noise in shaping genome architecture more broadly, it will be imperative to obtain comparative measures of noise and dosage sensitivity. In addition, we anticipate that the study of genome organization in three-dimensional space, powered by high-resolution chromatin capture techniques, will provide critical insights into adaptive interactions between genome architecture and expression processes. In particular, it will be interesting to explore whether spatial segregation of transcription foci within the nucleus serves to limit erroneous interactions, for example by confining promiscuously binding transcription factors to a defined nuclear domain.

In relation to gene anatomy we highlighted several cases where evidence strongly points towards selection having acted on transcript structure and composition to reduce the fitness burden of erroneous gene products. For many stages of gene expression, however, we remain ignorant about whether there are sequence-level adaptations to facilitate error prevention or mitigation. To advance our understanding in this regard, it will be critical to combine system-level molecular interaction data with knowledge about the evolutionary regime that governs fixation probabilities (Box 2) and the structure of pleiotropy in the system (Box 3).

Finally, as modern sequencing technologies continue to unearth increasingly more and rare isoforms, understanding molecular signatures of error adaptation may yield valuable clues for understanding what is functional diversity and what is not.

Tobias Warnecke; Centre for Genomic Regulation (CRG) and UPF; Carrer Dr. Aiguader, 88; 08003 Barcelona, Spain.

Laurence D Hurst; Department of Biology and Biochemistry, University of Bath; Claverton Down, BA27AY Bath, United Kingdom.

Correspondence to TW. e-mail: tobias.warnecke@crg.eu

or

LDH. e-mail: bssldh@bath.ac.uk

Box 1 | Telling apart functional and aberrant isoforms – the hard problem

With deep-sequencing platforms providing ample raw material for analyzing erroneous gene expression, a major challenge is to discriminate aberrant from functional transcript isoforms. Three criteria are frequently used to assess likely functionality: rarity, evolutionary conservation, and telltale sequence features. All, certainly when deployed individually, have their drawbacks.

Rarity. Very rare isoforms are sometimes assumed to be erroneously produced[20]. Although probably a good first-pass approximation, there are obvious pitfalls to equating erroneous with rare isoforms. Many transcripts with critical biological functions, including many transcription factors, are present in (very) few copies. Conversely, not everything that is common need be functional. If error mitigation is metabolically cheap and errors not immediately deleterious (i.e. cells can tolerate the erroneous product hanging around), erroneous isoforms might be much more ubiquitous than commonly assumed.

Conservation. Conservation of transcripts across species has also been used to categorize isoforms according to likely functionality. Many splicing isoforms are poorly conserved, bolstering the argument that mis-splicing is widespread[46,47]. Inevitably, this approach will yield some false positives, i.e. isoforms that look like errors but are not; these isoforms, being species-specific, might be particularly interesting to understand phenotypic variation between species. Conversely, there will be a number of false negatives: isoforms that do not look like errors but are. If error-prone states are maintained by pleiotropy (see Box 3), errors might be frequent, systematic, and systematically conserved. What the rate of such false negatives might be is completely unknown.

Telltale signs. For some expression processes there are telltale signs thought to indicate that the isoform was produced in error. The presence of premature termination codons (PTCs) is regarded as a strong indicator that something has gone awry. Yet even PTC-containing isoforms cannot be automatically

classified as errors. NMD-targeted isoforms of SR protein genes, for example, are highly conserved across mammals and participate in auto-regulatory feedback loops[48] making their production functional rather than erroneous. For primary transcripts, comparing its sequence to the DNA template can reveal the presence of transcription errors. However, observed discrepancies might largely be technical (i.e. sequencing errors) and post-transcriptional modifications (RNA editing) need to be ruled out.

Box 2 | The evolutionary context of error-proofing

In order to understand why some genes exhibit error-adaptive features yet others do not, we need to take into account the population genetic context in which genes and genomes have evolved.

In particular, the leverage of selection can vary substantially within and between genomes. Within genomes, other factors being equal, selection is stronger for more highly expressed genes, which accounts for the strong link between expression level and the degree of optimal codon usage as well as splicing fidelity[2, 20].

Between genomes, differences in effective population size (N_e) will affect the leverage of selection. The dearth of optimal codons in obligate endosymbionts, for example, is commonly attributed to stronger drift due to small population size[49]. Similarly, recent evidence suggests that average protein stability is reduced in small populations[50]. In the short term, this implies a higher error load owing to elevated rates of protein unfolding and undesirable interactions following the exposure of hydrophobic surfaces. In the long term, however, increased interactivity might facilitate the evolution of a more complex and versatile protein interactome[50].

Does reduced N_e therefore inevitably compromise the capacity to mitigate errors? Not necessarily. Although selection may become too weak to promote adaptation of individual genes, selection on system-wide mitigation mechanisms such as chaperones and NMD proteins may actually strengthen in line with the elevated workload from multiple increasingly poorly adapted substrates. Indeed, the prevalence, in small populations, of global versus local solutions to error mitigation was recently predicted by evolutionary modeling[51] and is consistent with the overexpression of GroEL in *Buchnera* and other endosymbionts (see Ref. 49 and references therein).

Beyond individual genes, the evolution of broad trends in genome architecture, such as genome size and the number of genes in the genome has also been attributed to differences in effective population size[52]. However, to what extent adaptive genome architecture such as the clustering of noise-sensitive genes breaks down under reduced $N_{e,,}$ remains largely unexplored.

Box 3 | Pleiotropy and the maintenance of error. Minimal error rates do not necessarily equate to optimal fitness, because reducing the incidence of errors usually comes at a cost. The textbook example here is translation where there is an intrinsic speed-accuracy trade-off that governs the interaction

between ribosome and mRNA[53,54]. Higher ribosomal accuracy is easily evolved but often selected against, because the resulting slow-down in protein production has a net negative effect on fitness[55]

A second type of trade-off concerns the coding potential of the information carrier (DNA, mRNA, etc.). Notably, protein-coding sequences not only specify amino acid content but also encode additional information about translational speed, secondary structure, and regulatory binding sites[56,57] . Error-prone sites may be maintained because a more accurate alternative causes a net fitness loss by compromising information unrelated to protein-coding, for example abrogating an exonic splicing enhancer[58].

The nature and severity of these trade-offs remains largely uncharacterized but should vary substantially across species depending, for example, on the growth strategy of the organism. The principal implication here is that high error levels may frequently be optimal, generating a persistent error reservoir even at large effective population sizes. We therefore speculate that, regardless of the capacity of selection to purge weakly deleterious mutations at individual sites (cf. Ref. 51), error mitigation might often be favoured during evolution. In turn, efficient error mitigation can dramatically lower the cost of gene expression errors and thereby alleviate the severity of pleiotropic trade-offs. This may have been an important factor in the evolution of regulatory complexity which – founded on combinatorial control – typically comes at a cost of making occasional errors [59, 60]. The presence of error mitigation can also permit rapid sampling of phenotypic space in processes such as V(D)J recombination, which generate a substantial number of non-functional and potentially harmful isoforms.

Figure 1. ***Error prevention and mitigation from transcription to protein folding*****.** At any point during gene expression, the relevant expression product is either error-free or has accumulated one or more errors. Both error-free and erroneous intermediates can progress further along the same processing stream or be degraded. In addition, new errors can be acquired and previous errors mitigated, for example when chaperones unfold or disaggregate a protein that had initially failed to fold correctly, so that error-free gene products can become erroneous and vice versa. These alternative processing fates, schematically depicted in the top half of the figure, have different consequences for fitness, with presumably beneficial and detrimental fates highlighted in green and red, respectively.

Features of the gene that promote error-free processing constitute adaptations for *error prevention.* Conversely, we can speak of *error mitigation* when an error has already occurred but that error is either corrected outright as in the chaperone example above (constructive mitigation) or its impact reduced, for example through targeting the erroneous transcript for degradation (destructive mitigation). The bottom half of the figure highlights some key steps in the expression of protein-coding genes along with examples of error prevention as well as constructive and destructive mitigation (see main text for details).

Figure 2. ***Common codons encode partial termination signals.*** A heat map illustrating codon usage frequencies in the human genome. When a one-base pair frame-shift occurs in either the upstream (-1) or

downstream (+1) direction, many codons form part of a stop codon in the new reading frame, which terminates translation. These codons are significantly more abundant on average than those that do not introduce a partial stop codon (P=0.025, logistic regression).

Acknowledgements

TW is the recipient of an EMBO Long-term Fellowship. LDH is a Royal Society Wolfson Research Merit Award Holder.

Competing interests statement

The authors declare that they have no competing financial interests.

Glossary

NONSENSE-MEDIATED DECAY

The process by which mRNAs containing premature termination codons are destroyed to preclude the production of truncated and potentially deleterious protein products. It is also used in combination with specific alternative splicing events to control the levels of some proteins

BURSTING

The pulse-like, non-continuous mode of transcript production where periods of active transcription are interspersed by periods of inactivity. Bursting may be a general feature of transcriptional activity and has been observed in both prokaryotic and eukaryotic cells.

BIOGRAPHIES

Tobias Warnecke received his BA in Human Sciences from the University of Oxford in 2006. Leaving one historic city for another, he moved to Laurence Hurst's lab at the University of Bath where he completed his PhD in 2010. After being awarded an EMBO Long-term Fellowship he joined Fyodor Kondrashov's group at the Centre for Genomic Regulation in Barcelona.

Laurence Hurst obtained his BA in Zoology from the University of Cambridge, UK, in 1987. After a one-year fellowship at Harvard, he did his DPhil on genetic conflicts under W.D. Hamilton and Alan Grafen at the University of Oxford, UK, completing in 1991. Upon receiving a Royal Society Research Fellowship he moved back to Cambridge, where he stayed for three years before being appointed Chair in Evolutionary Genetics at the University of Bath. He is the holder of a Royal Society Wolfson

Research Merit Award. He is interested in fundamental problems concerning the evolution of genes, genomes and genetic systems.

Author websites:

Tobias Warnecke: http://big.crg.cat/people/twarnecke

Laurence Hurst: http://people.bath.ac.uk/bssldh/LaurenceDHurst/Home.html

References

Text for table of contents

Drawing on evidence from diverse species, the authors argue that aspects of gene anatomy and genome architecture have evolved to prevent or mitigate gene expression errors. These aspects range from codon usage to the arrangement of genes in the genome.

1. Fox-Walsh, K.L. & Hertel, K.J. Splice-site pairing is an intrinsically high fidelity process. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 1766-1771 (2009).
2. Drummond, D.A. & Wilke, C.O. The evolutionary consequences of erroneous protein synthesis. *Nat. Rev. Genet.***10**, 715-724 (2009).
3. Ackermann, M. & Chao, L. DNA Sequences shaped by selection for stability. *PLoS Genet.* **2**, e22 (2006).
4. Wagner, L.A., Weiss, R.B., Driscoll, R., Dunn, D.S. & Gesteland, R.F. Transcriptional slippage occurs during elongation at runs of adenine or thymine in Escherichia coli. *Nucleic Acids Res.* **18**, 3529-3535 (1990).
5. Weiss, R.B., Dunn, D.M., Atkins, J.F. & Gesteland, R.F. Slippery runs, shifty stops, backward steps, and forward hops: -2, -1, +1, +2, +5, and +6 ribosomal frameshifting. *Cold Spring Harb. Symp. Quant. Biol.* **52**, 687-693 (1987).
6. Woese, C.R. On the evolution of the genetic code. *Proc. Natl. Acad. Sci. U. S. A.* **54**, 1546-52 (1965).
7. Massey, S.E. A neutral origin for error minimization in the genetic code. *J. Mol. Evol.* **67**, 510-516 (2008).
8. Freeland, S.J. & Hurst, L.D. The genetic code is one in a million. *J. Mol. Evol.* **47**, 238-248 (1998).
9. Khajavi, M., Inoue, K. & Lupski, J.R. Nonsense-mediated mRNA decay modulates clinical outcome of genetic disease. *Eur. J. Hum. Genet.* **14**, 1074-1081 (2006).
10. Maquat, L.E. & Carmichael, G.G. Quality control of mRNA function. *Cell* **104**, 173-176 (2001).
11. Jaillon, O. et al. Translational control of intron splicing in eukaryotes. *Nature* **451**, 359-362 (2008).
12. Itzkovitz, S. & Alon, U. The genetic code is nearly optimal for allowing additional information within protein-coding sequences. *Genome Res.* **17**, 405-412 (2007).
13. Mekouar, M. et al. Detection and analysis of alternative splicing in Yarrowia lipolytica reveal structural constraints facilitating nonsense-mediated decay of intron-retaining transcripts. *Genome Biol.* **11** (2010).
14. Seligmann, H. & Pollock, D.D. The ambush hypothesis: hidden stop codons prevent off-frame gene reading. *DNA Cell Biol.* **23**, 701-705 (2004).
15. Warnecke, T., Huang, Y., Przytycka, T.M. & Hurst, L.D. Unique cost dynamics elucidate the role of frameshifting errors in promoting translational robustness. *Genome Biol. Evol.***2**, 636-645 (2010).
16. Clarke, C.H. The consequences of base-pair substitution mutations in AT- and GC-rich bacteria. *J. Theor. Biol.* **105**, 117-131 (1983).

17. Cusack, B.P., Arndt, P.F., Duret, L. & Crollius, H.R. Preventing dangerous nonsense: selection for robustness to transcriptional error in human genes. *PLoS Genet.* **7**, e1002276 (2011).
18. Warnecke, T. & Hurst, L.D. GroEL dependency affects codon usage-support for a critical role of misfolding in gene evolution. *Mol. Syst. Biol.* **6** (2010).
19. Kaida, D. et al. U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation. *Nature* **468**, 664-668 (2010).
20. Pickrell, J.K., Pai, A.A., Gilad, Y. & Pritchard, J.K. Noisy splicing drives mrna isoform diversity in human cells. *PLoS Genet.* **6** (2010).
21. Frischmeyer, P.A. et al. An mRNA surveillance mechanism that eliminates transcripts lacking termination codons. *Science* **295**, 2258-2261 (2002).
22. Ito-Harashima, S., Kuroha, K., Tatematsu, T. & Inada, T. Translation of the poly(A) tail plays crucial roles in nonstop mRNA surveillance via translation repression and protein destabilization by proteasome in yeast. *Gen. Dev.* **21**, 519-524 (2007).
23. Ito, K. et al. RUNX3, a novel tumor suppressor, is frequently inactivated in gastric cancer by protein mislocalization. *Cancer Res.* **65**, 7743-7750 (2005).
24. Hilleren, P., McCarthy, T., Rosbash, M., Parker, R. & Jensen, T.H. Quality control of mRNA 3'-end processing is linked to the nuclear exosome. *Nature* **413**, 538-542 (2001).
25. Hurst, L.D., Pál, C. & Lercher, M.J. The evolutionary dynamics of eukaryotic gene order. *Nat. Rev. Genet.***5**, 299-310 (2004).
26. Rocha, E.P. & Danchin, A. Essentiality, not expressiveness, drives gene-strand bias in bacteria. *Nat. Genet.* **34**, 377-378 (2003).
27. Raser, J.M. & O'Shea, E.K. Noise in gene expression: origins, consequences, and control. *Science* **309**, 2010-3 (2005).
28. Newman, J.R. et al. Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise. *Nature* **441**, 840-846 (2006).
29. Wang, Z. & Zhang, J. PNAS Plus: Impact of gene expression noise on organismal fitness and the efficacy of natural selection. *Proc. Natl. Acad. Sci. U. S. A.* **108**, E67-76 (2011).
30. Woo, Y.H. & Li, W.H. Gene clustering pattern, promoter architecture, and gene expression stability in eukaryotic genomes. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 3306-3311 (2011) .
31. Tirosh, I. & Barkai, N. Two strategies for gene regulation by promoter nucleosomes. *Genome Res.* **18**, 1084-1091 (2008).
32. Becskei, A., Kaufmann, B.B. & van Oudenaarden, A. Contributions of low molecule number and chromosomal positioning to stochastic gene expression. *Nat. Genet.* **37**, 937-944 (2005).
33. Becskei, A. & Serrano, L. Engineering stability in gene networks by autoregulation. *Nature* **405**, 590-593 (2000).
34. Kollmann, M., Løvdok, L., Bartholomé, K., Timmer, J. & Sourjik, V. Design principles of a bacterial signalling network. *Nature* **438**, 504-507 (2005).
35. Batada, N.N. & Hurst, L.D. Evolution of chromosome organization driven by selection for reduced gene expression noise. *Nat. Genet.* **39**, 945-949 (2007).
36. Raj, A., Peskin, C.S., Tranchina, D., Vargas, D.Y. & Tyagi, S. Stochastic mRNA synthesis in mammalian cells. *Plos Biol.* **4**, e309 (2006).
37. Wang, G.Z., Lercher, M.J. & Hurst, L.D. Transcriptional coupling of neighboring genes and gene expression noise: evidence that gene orientation and noncoding transcripts are modulators of noise. *Genome Biol. Evol.Genome Biol. Evol.* **3**, 320-331 (2011).
38. Xu, Z. et al. Antisense expression increases gene expression variability and locus interdependency. *Mol. Syst. Biol.* **7** (2011).
39. Lovdok, L. et al. Role of translational coupling in robustness of bacterial chemotaxis pathway. *Plos Biol.* **7** (2009).
40. Kovacs, K., Hurst, L.D. & Papp, B. Stochasticity in protein levels drives colinearity of gene order in metabolic operons of Escherichia coli. *Plos Biol.* **7** (2009).
41. Kamath, R.S. et al. Systematic functional analysis of the Caenorhabditis elegans genome using RNAi. *Nature* **421**, 231-237 (2003).
42. de Clare, M., Pir, P. & Oliver, S.G. Haploinsufficiency and the sex chromosomes from yeasts to humans. *BMC Biol.* **9**, 15 (2011).
43. Cook, D.L., Gerber, A.N. & Tapscott, S.J. Modeling stochastic gene expression: implications for haploinsufficiency. *Proc. Natl. Acad. Sci. U.S.A* **95**, 15641-15646 (1998).
44. Yin, S.Y. et al. Dosage compensation on the active X chromosome minimizes transcriptional noise of X-linked genes in mammals. *Genome Biol.***10** (2009).

45. Hegarty, M.J. et al. Transcriptome shock after interspecific hybridization in *senecio* is ameliorated by genome duplication. *Curr. Biol.* **16**, 1652-1659 (2006).
46. Melamud, E. & Moult, J. Stochastic noise in splicing machinery. *Nucleic Acids Res.* **37**, 4873-4886 (2009).
47. Tress, M.L. et al. The implications of alternative splicing in the ENCODE protein complement. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 5495-5500 (2007).
48. Lareau, L., Brooks, A., Soergel, D., Meng, Q. & Brenner, S. in Alternative splicing in the postgenomic era (eds. Blencowe, B. & Graveley, B.) 191-212 (Landes Biosciences, Austin, TX, 2007).
49. Wernegreen, J.J. & Moran, N.A. Evidence for genetic drift in endosymbionts (Buchnera): analyses of protein-coding genes. *Mol. Biol. Evol.* **16**, 83-97 (1999).
50. Fernández, A. & Lynch, M. Non-adaptive origins of interactome complexity. *Nature* **474**, 502-505 (2011).
51. Rajon, E. & Masel, J. Evolution of molecular error rates and the consequences for evolvability. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 1082-1087 (2011).
52. Lynch, M. The origins of genome architecture (Sinauer Associates, Sunderland, Mass., 2007).
53. Thompson, R.C. & Karim, A.M. The accuracy of protein biosynthesis is limited by its speed: high fidelity selection by ribosomes of aminoacyl-tRNA ternary complexes containing GTP[gamma S]. *Proc. Natl. Acad. Sci. U. S. A.* **79**, 4922-4926 (1982).
54. Wohlgemuth, I., Pohl, C. & Rodnina, M.V. Optimization of speed and accuracy of decoding in translation. *EMBO J.* **29**, 3701-3709 (2010).
55. Ruusala, T., Andersson, D., Ehrenberg, M. & Kurland, C.G. Hyper-accurate ribosomes inhibit growth. *EMBO J.* **3**, 2575-2580 (1984).
56. Itzkovitz, S., Hodis, E. & Segal, E. Overlapping codes within protein-coding sequences. *Genome Res.* **20**, 1582-1589 (2010).
57. Warnecke, T., Weber, C.C. & Hurst, L.D. Why there is more to protein evolution than protein function: splicing, nucleosomes and dual-coding sequence. *Biochem. Soc. Trans.* **37**, 756-761 (2009).
58. Warnecke, T. & Hurst, L.D. Evidence for a trade-off between translational efficiency and splicing regulation in determining synonymous codon usage in Drosophila melanogaster. *Mol. Biol. Evol.* **24**, 2755-2762 (2007).
59. Boue, S., Letunic, I. & Bork, P. Alternative splicing and evolution. *Bioessays* **25**, 1031-1034 (2003).
60. Doma, M.K. & Parker, R. RNA quality control in eukaryotes. *Cell* **131**, 660-668 (2007).

Figure 1

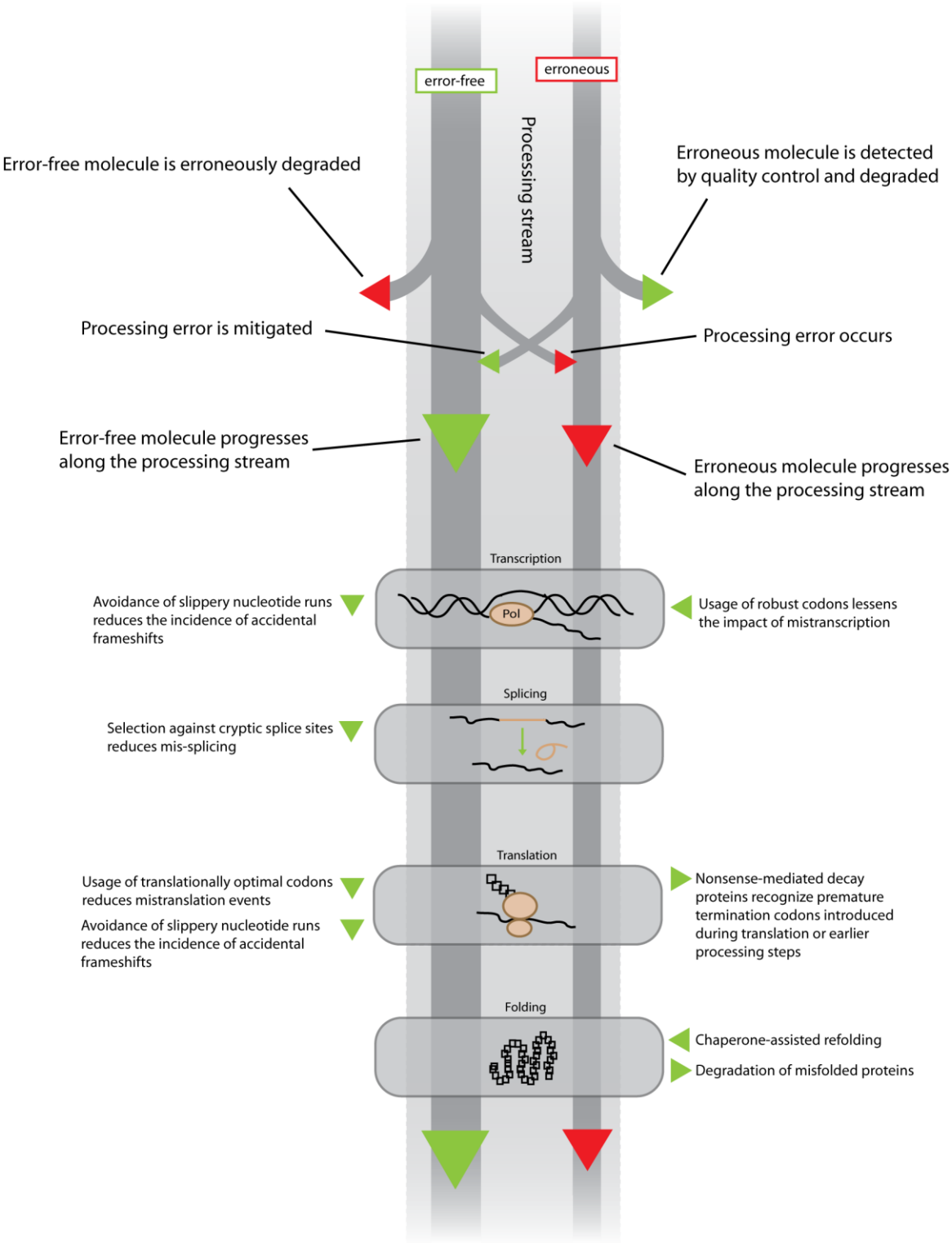
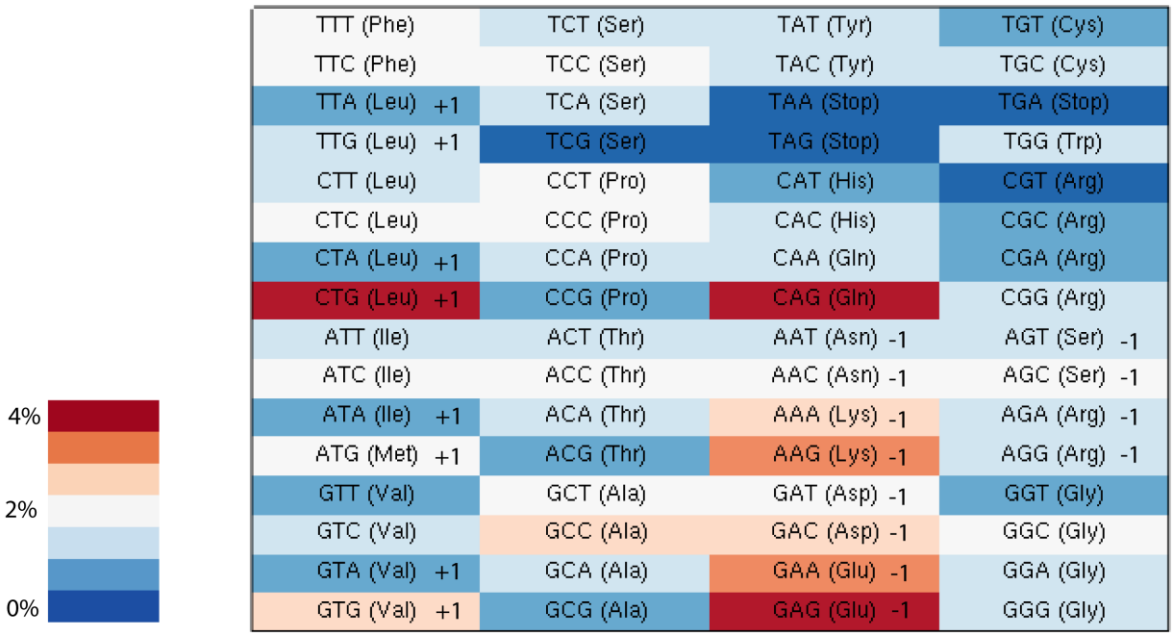

error-free
erroneous
Processing stream
Error-free molecule is erroneously degraded
Erroneous molecule is detected by quality control and degraded
Processing error is mitigated
Processing error occurs
Error-free molecule progresses along the processing stream
Erroneous molecule progresses along the processing stream
Transcription
Pol
Avoidance of slippery nucleotide runs reduces the incidence of accidental frameshifts
Usage of robust codons lessens the impact of mistranscription
Splicing
Selection against cryptic splice sites reduces mis-splicing
Translation
Usage of translationally optimal codons reduces mistranslation events
Avoidance of slippery nucleotide runs reduces the incidence of accidental frameshifts
Nonsense-mediated decay proteins recognize premature termination codons introduced during translation or earlier processing steps
Folding
Chaperone-assisted refolding
Degradation of misfolded proteins

Figure 2


4%
2%
0%
TTT (Phe)
TTC (Phe)
TTA (Leu) +1
TTG (Leu) +1
CTT (Leu)
CTC (Leu)
CTA (Leu) +1
CTG (Leu) +1
ATT (Ile)
ATC (Ile)
ATA (Ile) +1
ATG (Met) +1
GTT (Val)
GTC (Val)
GTA (Val) +1
GTG (Val) +1
TCT (Ser)
TCC (Ser)
TCA (Ser)
TCG (Ser)
CCT (Pro)
CCC (Pro)
CCA (Pro)
CCG (Pro)
ACT (Thr)
ACC (Thr)
ACA (Thr)
ACG (Thr)
GCT (Ala)
GCC (Ala)
GCA (Ala)
GCG (Ala)
TAT (Tyr)
TAC (Tyr)
TAA (Stop)
TAG (Stop)
CAT (His)
CAC (His)
CAA (Gln)
CAG (Gln)
AAT (Asn) -1
AAC (Asn) -1
AAA (Lys) -1
AAG (Lys) -1
GAT (Asp) -1
GAC (Asp) -1
GAA (Glu) -1
GAG (Glu) -1
TGT (Cys)
TGC (Cys)
TGA (Stop)
TGG (Trp)
CGT (Arg)
CGC (Arg)
CGA (Arg)
CGG (Arg)
AGT (Ser) -1
AGC (Ser) -1
AGA (Arg) -1
AGG (Arg) -1
GGT (Gly)
GGC (Gly)
GGA (Gly)
GGG (Gly)